Phân Tích Dữ Liệu Thực Tế với Python

Bài 8.1: Phân Tích Dự Đoán

Quang-Khai Tran, Ph.D
CyberLab, 03/2023

(Ảnh: Internet)

Nội dung

1. Giới thiệu
2. Linear Regression
3. Logistic Regression
4. Bài tập & Thảo Luận

Phần 1. Giới thiệu

Phân tích dự đoán/dự báo là gì?

1 Phân Nhóm Các Loại dữ liệu trong thống kê

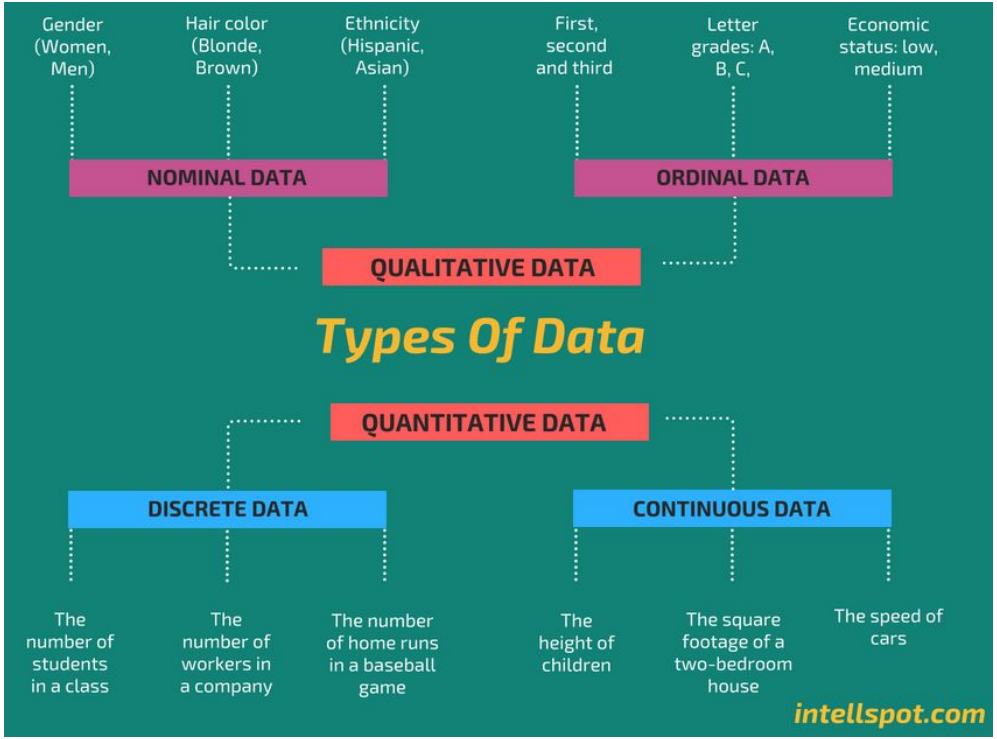
Quantitative (định lượng):
- discrete (rời rạc)
- continuous (liên tục)
- interval (khoảng)

  ⇒ structured data

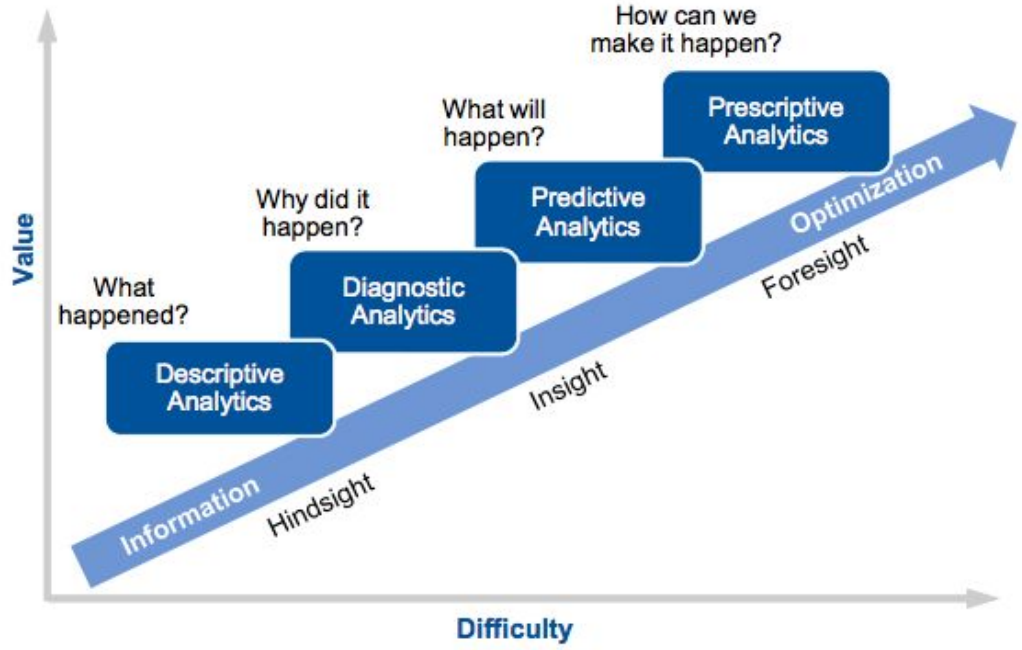
Qualitative (định tính):
- nominal (định danh)
- binary (định danh True/False)
- ordinal (thứ tự)

  ⇒ unstructured data

(text, phân loại, datetime)



Source: https://www.intellspot.com/data-types/



Source: Gartner (March 2012)

1 Phân tích dự đoán/dự báo

- Phân tích dự báo/dự đoán định lượng:
  - Dự đoán dựa vào tương quan và hồi quy (prediction/forecast)
    - Ước lượng giá trị (regression)
    - Ước lượng phân loại (classification)
    - Kết hợp ước lượng giá trị và phân loại
  - Dự đoán đối tượng ngoại biên (phân tích sự bất thường, outliers prediction)
  - Dự đoán dữ liệu chuỗi thời gian (time series forecasting):
    - Biến đổi theo mùa (seasonality)
    - Biến đổi theo chu kỳ (cycles)
    - Biến đổi ngẫu nhiên (random variations)
- Phân tích sự phân cụm (clustering)

1 Phân tích dự đoán/dự báo

❖ 5 loại mô hình phân tích dự báo (theo Oracle)

ORACLE NETSUITE

1. Mô hình dự báo/dự đoán (prediction/forecast)	Dự đoán một giá trị trong tương lai dựa vào dữ liệu quá khứ (linear/non-linear)
2. Mô hình phân loại (classification)	Xác định loại của từng sample dữ liệu
3. Mô hình ngoại biên (outliers)	Tìm ra/dự đoán đối tượng bất thường
4. Mô hình chuỗi thời gian (time series)	Đánh giá một chuỗi các dữ liệu thời gian
5. Mô hình phân cụm (clustering)	Phân nhóm các dữ liệu dựa theo đặc điểm chung của chúng

Source: https://www.netsuite.com/portal/resource/articles/financial-management/predictive-modeling.shtml

Phân tích dự đoán/dự báo

❖ 3 cách mô hình hóa phân tích dự báo (theo IBM)

1. Mô hình hóa tính dự đoán (predictive modeling)	Sử dụng thống kê, AI để dự đoán outcomes (các giá trị muốn đạt được)
2. Mô hình hóa tính mô tả (descriptive modeling)	Mô tả, khám phá ra các mối quan hệ trong dữ liệu nhằm phục vụ dự đoán
3. Mô hình hóa quyết định (decision modeling)	Mô tả, khám phá ra các mối quan hệ giữa các thành phần của quyết định, ảnh hưởng của quyết định đến các thành phần

Source: https://www.ibm.com/analytics/predictive-analytics

Một số lợi ích:

- Giảm thời gian, công sức và chi phí để dự báo outcomes
  - Các yếu tố thuộc về môi trường kinh doanh
  - Thay đổi về quy định
  - Đối thủ
  - Các điều kiện thị trường

- Phục vụ hiệu quả cho việc ra quyết định, phân tích đề nghị

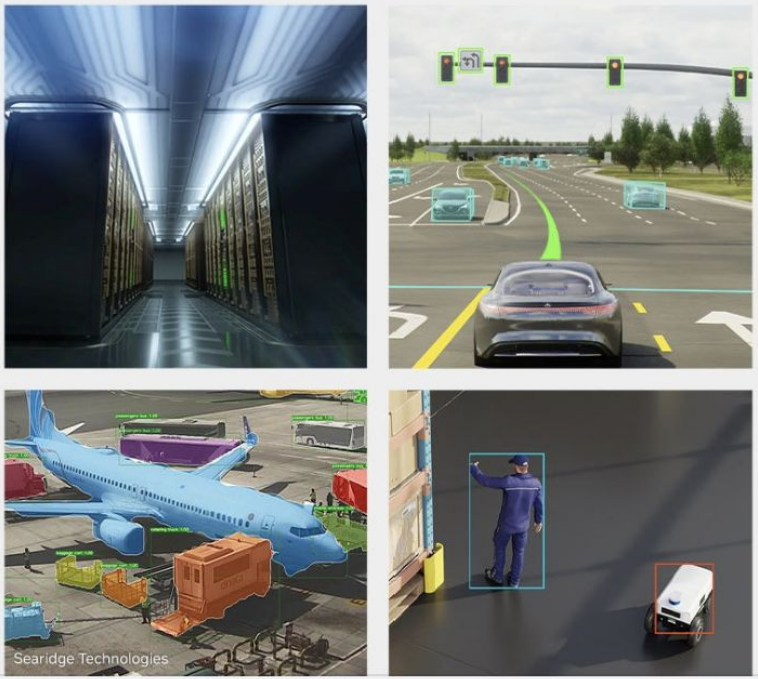
Phân tích dự đoán/dự báo

Một số ví dụ:

- Tài chính - chứng khoán
  - Dự đoán giá chứng khoán, tăng vs. giảm, triển vọng
  - Dự đoán khách hàng rời bỏ (churn prediction)
  - Tự động phát hiện các giao dịch gian lận, ước lượng sự đáng ngờ

- Bán lẻ - Dịch vụ:
  - Dự đoán doanh thu, doanh số, số lượng khách hàng
  - Dự đoán thời gian giao hàng
  - Phân loại khách hàng, sản phẩm, dịch vụ

The Technology Conference for the Era of AI and the Metaverse

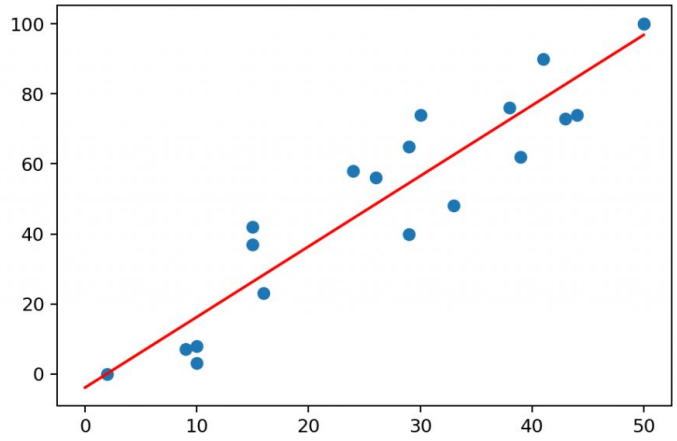


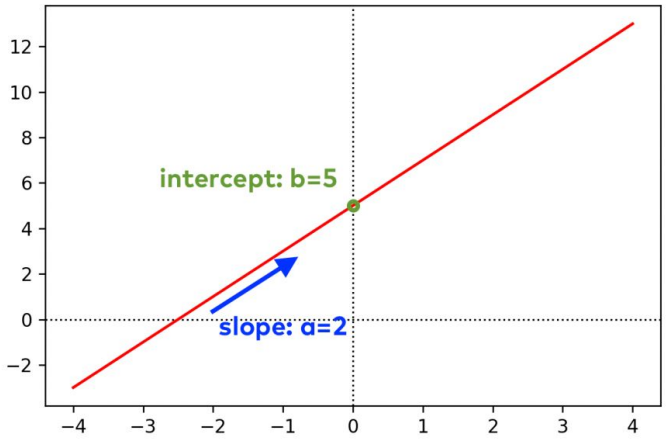
Phần 2.
Linear Regression

2.1 Giới thiệu Linear Regression

Hồi quy tuyến tính

- Là giải thuật dự đoán giá trị của dữ liệu dựa vào sự tương quan tuyến tính
- Hàm: y = ax + b





Các hạn chế

- Không bao giờ tìm được "perfect fit"
- Khả năng dự đoán "tốt" chỉ giới hạn ở dữ liệu đã có
- Không giải thích được (tường tận) cách thức các biến tương tác với nhau

2.1 Một số thư viện/hàm hỗ trợ trong Python

Numpy:

```
np.polyfit()
np.poly1d((a,b))
```

Scipy:

```
slope, intercept, r, p, std_err = stats.linregress(x, y)
```

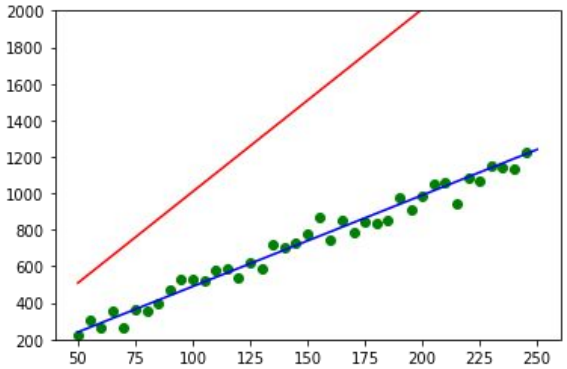
Sklearn:

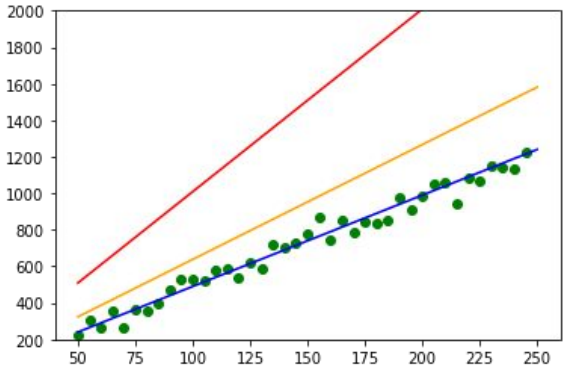
```
from sklearn.linear_model import LinearRegression
```

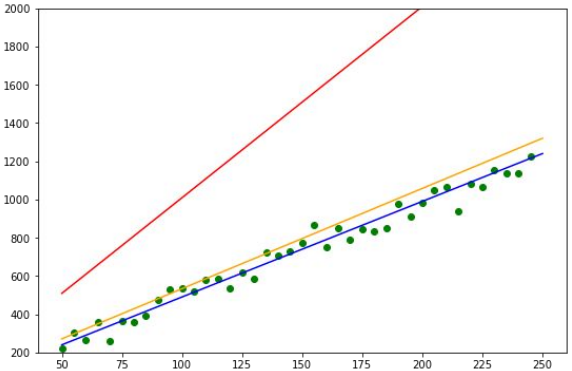
StatsModels:

```
import statsmodels.api as sm
model = sm.OLS(Y, X).fit()
predictions = model.predict(X)
```

1. Demo-1: bản chất của linear regression với phương trình y=ax+b (Hồi quy tuyến tính đơn biến)







Demo-2: sử dụng hồi quy tuyến tính đa biến để dự đoán giá nhà (Boston housing price dataset)

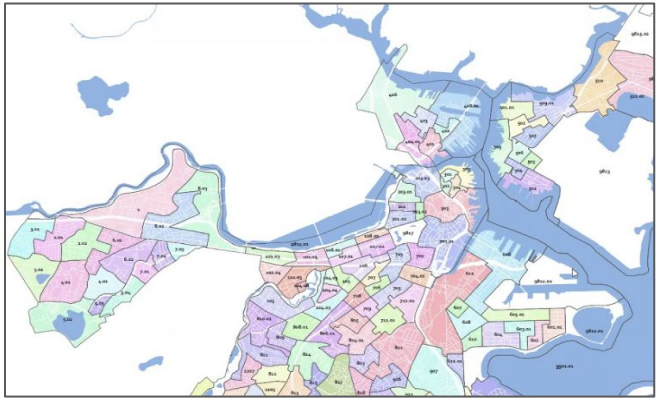
```
------
Data Set Characteristics:

    :Number of Instances: 506

    :Number of Attributes: 13 numeric/categorical predictive

    :Median Value (attribute 14) is usually the target

    :Attribute Information (in order):
        - CRIM     per capita crime rate by town
        - ZN       proportion of residential land zoned for lots over 25,000 sq.ft.
        - INDUS    proportion of non-retail business acres per town
        - CHAS     Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
        - NOX      nitric oxides concentration (parts per 10 million)
        - RM       average number of rooms per dwelling
        - AGE      proportion of owner-occupied units built prior to 1940
        - DIS      weighted distances to five Boston employment centres
        - RAD      index of accessibility to radial highways
        - TAX      full-value property-tax rate per $10,000
        - PTRATIO  pupil-teacher ratio by town
        - B        1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
        - LSTAT    % lower status of the population
        - MEDV     Median value of owner-occupied homes in $1000's
```

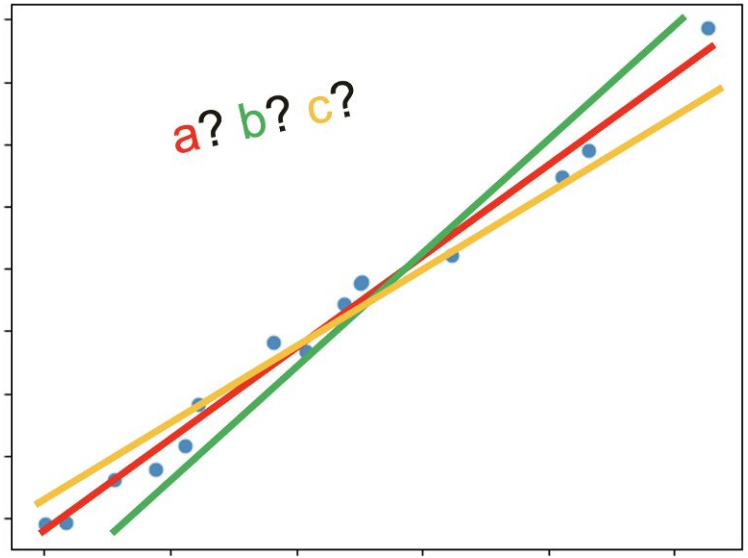


Source: https://www.weirdgeek.com/2018/12/linear-regression-to-boston-housing-dataset/

2.2 Hồi quy tuyến tính đơn biến

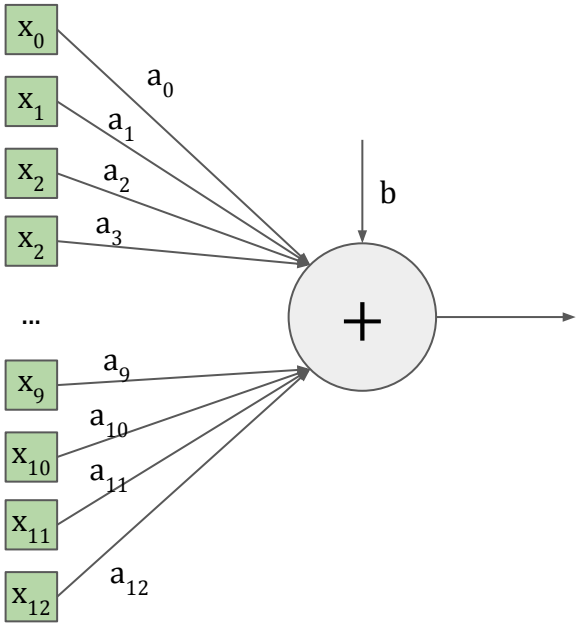
Univariate Linear Regression:

Biến output chỉ phụ thuộc vào một biến input

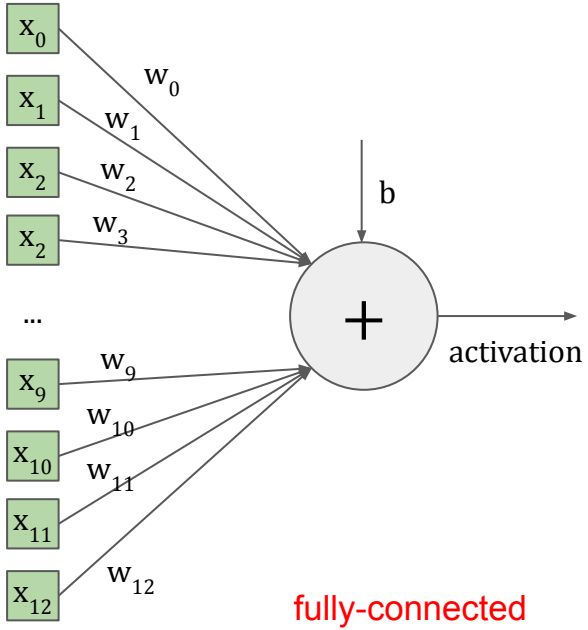


2.3 Hồi quy tuyến tính đa biến

Multivariate Linear Regression



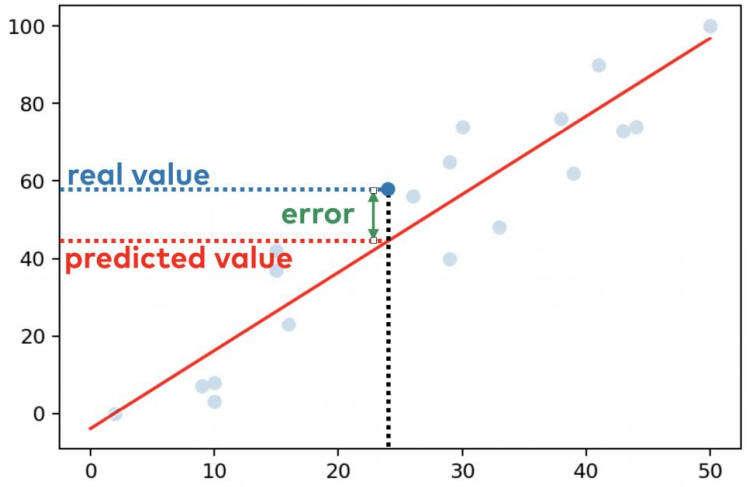
Simplest neural network

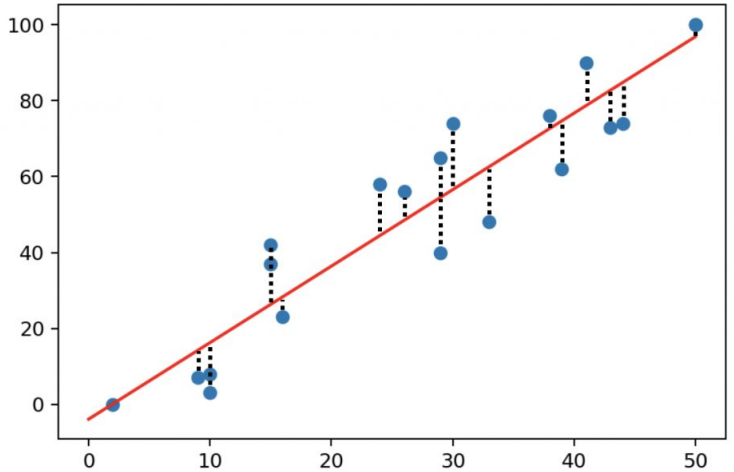


fully-connected

2.4 Cách đánh giá

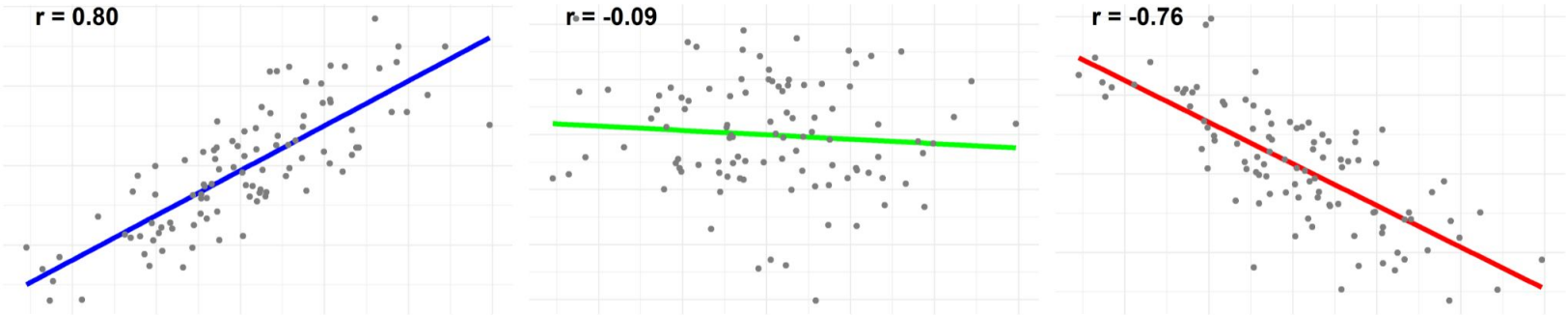
- Một số lỗi để đánh giá mức độ dự đoán không chính xác của mô hình
  - Error: predicted - groundtruth
  - MSE, RMSE, MAE





2.4 Cách đánh giá

- Correlation coefficient (R)
- Coefficient of determination (R-squared)



2.4 Cách đánh giá

- Ôn lại: ***hệ số tương quan R*** (Pearson's correlation coefficient)
  - Đo lường mức độ của mối quan hệ tuyến tính giữa 02 (hai) biến x_i và x_j
  - Khoảng giá trị: $-1 \leq R \leq +1$

⇒ Lưu ý: không dùng R để đo lường độ chính xác của mô hình dự đoán
⇒ Nên dùng để xác định các biến nào tốt cho việc dự đoán

$$\rho_{xy} = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

2.4 Cách đánh giá

❖ R-squared (coefficient of determination: **hệ số xác định**,

Lưu ý: một số tài liệu gọi là **hệ số tương quan bội của số R-bình phương**)

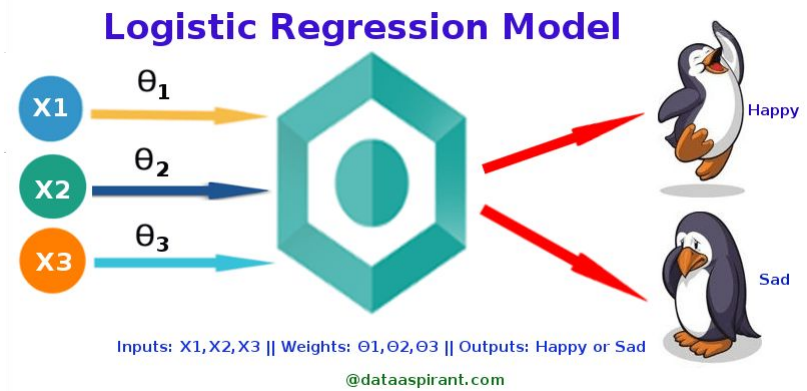
- "Tỷ lệ (hoặc phần trăm) của sự thay đổi (proportion/percentage of variation) trong biến phục thuộc được dự đoán từ các biến độc lập" (***theo wikipedia***)
- Ví dụ: $y = a_1x_1 + a_2x_2 + b$, và $R^2 = 0.88$
  $\Rightarrow x_1$ và x_2 giải thích 88% sự thay đổi của y
- Được sử dụng để đo lường khả năng một mô hình linear regression "khớp vào" (fitting) dữ liệu (***the goodness of fitting***)
- Khoảng giá trị: $0 \leq R^2 \leq 1$ (trông đợi: $R^2 >= 0.81$)
  - Giá trị cao: khả năng dự đoán tốt của các biến độc lập được chọn
  - Giá trị bằng 1: khả năng dự đoán là chính xác
- Training: R^2; Testing: Q^2 (***the goodness of predicting***)

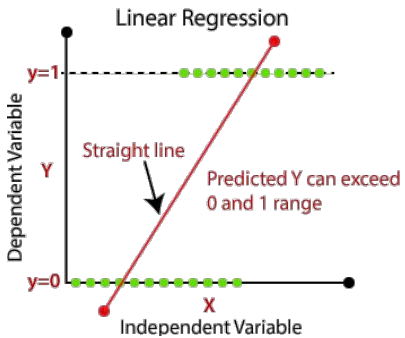
Phần 3.
Logistic Regression

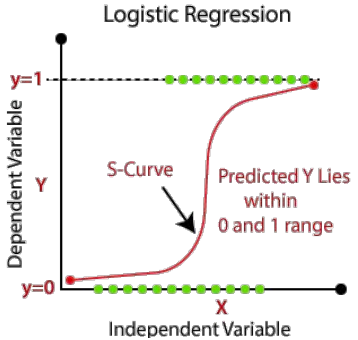
3.1 Giới thiệu về Logistic Regression

Hồi quy logistic

- Nhằm dự đoán giá trị đầu ra rời rạc (discrete target variables)
- Tương tự như phân loại các đầu vào x vào các nhóm y tương ứng







3.1 Giới thiệu về Logistic Regression

Hồi quy logistic

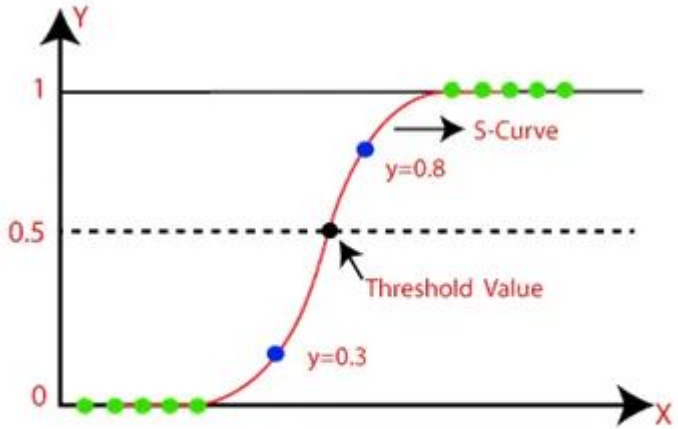
- Dựa vào hàm logarit
- Để phân loại: hàm sigmoid

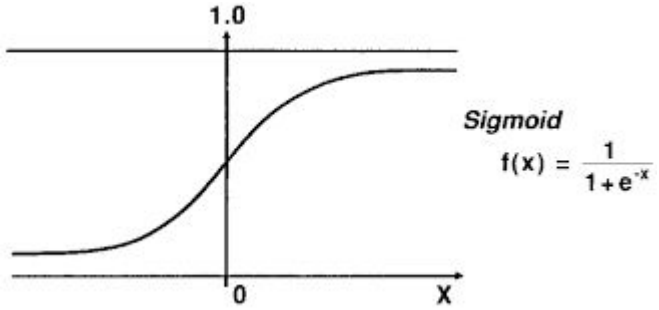
a.k.a. **Log Odds**

or **Logit**

Intercept

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X$$





Sử dụng Logistic regression trong Python với thư viện `sklearn`

```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(x_train, y_train)
y_hat = regressor.predict(x_test)
```

3.2 Cách đánh giá

- Accuracy: 100*(no. of correct predictions / total no. of predictions)
- Error: 1 - accuracy
- Confidence Interval:
  (Note: assume that n > 30)
  - Error: interval = z * sqrt((error * (1 - error)) / n)
  - Accuracy: interval = z * sqrt((accuracy * (1 - accuracy)) / n)

  (90%: z=1.64; 95%: z=1.96; 98%: z=2.33; 99%: z=2.58)

Demo-3: sử dụng logistic regression để dự đoán chất lượng xe hơi cũ

4 Bài tập

Thực hiện phân tích mô tả cho dữ liệu chuyến bay "nycflights.csv":

1. Cho cột dep_delay (khởi hành trễ)
2. Cho cột arr_delay (đến nơi trễ)
3. Cho cột distance (khoảng cách chuyến bay)
4. Sử dụng linear regression xây dựng mô hình dự đoán thời gian đến nơi trễ (arr_delay) dựa vào thời gian xuất phát trễ (dep_delay) và khoảng cách (distance).

Bài tập

Nâng cao: tiếp tục với dữ liệu 'nycflights.csv'

1. Thực hiện phân tích mô tả cho 3 cột ở trên nhưng chia theo nơi xuất phát: cột origin (gồm 3 sân bay: JFK, LGA, EWR)
2. Chia ra làm 3 mô hình cho 3 sân bay xuất phát (JFK, LGA, EWR) và nhận xét về độ chính xác so với mô hình chung

THANK YOU!